Deep and Wide Model in Automated Essay Scoring

Luyao Peng

2020-01-17

## Introduction

Generic approach in automated essay scoring produces scores that have the same standards/meanings across all prompts, existing or new, of a writing assessment. This is accomplished by using a single set of indicators or features, a consistent way of weighting and transforming these features into essay scores, and a focus on features that are not based on prompt-specific information or vocabulary.

Tranditional generic approach focuses on creating linguistic features that are designed to measure elements of writing quality and combining and weighting those features to produce an essay score by using machine learning methods (current CRASE+ engine applies gradient-boosting regression method). The advantages of using traditional machine learning methods are 1. it is suitable machine learning problems with small sample sizes (which is currently true for automated essay scoring field); 2. it has good iterpretibility about not only the meaningful features but also how the features are combined and weighted to produce an essay score, which we call that it has good memorization (in memorizing the relationships and effects between input features and output score). However, it is also because that the traditional machine learning methods have good and relatively fixed memorization of the input and output in the training essays, it somewhat loses the flexibility of generalizing the scoring rules to essays from a new prompt.

Building generic model requires the model to be able to generalize scoring standards and reliability to the essays from new prompts (the prompts that have not being trained). In recent years, learning with deep neural networks has enjoyed huge empirical success across a widevariety of tasks. More surprisingly, the networks learned this way exhibit good generalization behavior due to the fact that deep learning is able to extract high-level abstractions of features. The common deep neural network in Natural Language Processing includes feed-forward neural network (FNN), recurrent neural network (RNN), long short term memory (LSTM) and transformer based neural network. As the number of essays in ACT writing assessments accumulate over time, deep learning methods will be benefited from the growing sample sizes and open a new resaerch direction in the field of automated essay scoring. However, one of the drawbacks of deep neural network is that it lacks of interpretability because the features being used in deep learning usually don’t carry interpritable meaning, the relation between features and essay scores, how essay features affect resulting scores are also unexplanable.

As we were exploring how to advance automated essay scoring, we asked ourselves the question—can we build a generic model with both good memorization and generalization, by combining the power of tradtional machine learning and deep learning methods?

The answer is ‘yes, we can’. So here comes the model, by jointly training a wide linear model (for memorization) alongside a deep neural network (for generalization), which was called the deep and wide learning and first proposed by Google in 2016. The deep and wide model has been studied in different areas such as ranking, search and recommendation system, but hasn’t been applied in automated essay scoring problem.

## The Architecture of Deep and Wide

The model is composed of two parts: deep model and wide model. Figure 1 demonstrated the architecture of deep and wide model being studied in CRASE team.

1. Deep: from bottom to the top are
2. the input of the deep model is the individual word in each essay;
3. representing each word using a dense vector of specific dimension (vectors are parameters to be learned during training);
4. taking average of the word vectors across all words in each essay;
5. two layers of LSTM with the number of output features being 128;
6. Wide: the features (sparse or dense) in the wide model include 40 CRASE+ designed linguistic features and normalized to be within [0, 1]

3 Put together: 1. concatenating 40 CRASE+ features in the wide model and 128 output features from the second layer of LSTM from the deep model results in 168 input features for prediction layer; 2. Fitting sigmoid function (logistic regression function) to the 168 features (by concatenating wide and deep features) and the essay score.

## Results

Deep and wide model was trained on 4 prompts from ACT International Writing assessment administed 2018-2019 (4136 essays) and tested on essays from remaining prompts (2044 essays from remaining 8 prompts). The results show that deep and wide method is competitive to the standard CRASE+ engine on certain metrics.

## Extensions

Even though the deep and wide learning was competative with the standard CRASE+ engine, it has the following drawbacks:

1. The high variance in results has been a well known fact about deep learning method, due to the randomization nature and intention of deep learning method. The unstablenss of deep learning’s results poses a question to use when scoring an essay: how confident are we in being certain that the quality of an essay is truly reflected by the predicted score?
2. Deep and wide model takes long time to train because it is well-known that the deep model has much larger number of parameters to be learned.
3. Including deep learning model usually requires larger number of training samples in order to achieve a better performances. This is the ‘data-hungry’ side of deep learning method.

Several extensions are being considered/studied to solve the drawbacks mentioned above:

1.GPDW:

To deal with high variation in results, ensemble learning is a common technique. In our current study, Gaussian process is considered to be combined with the deep and wide model, the resulting model is called Gaussian Process Deep and Wide model (GPDW), which may be helpful in reducing the high variation and overfitting problems of standard deep and wide model, and it can be shown mathematically the equivalency between GPDW and the ensemble of infinite number of neural network.

1. GPDW is essentailly providing essay score predictions based on posterior conditional distribution of the essay score given known essay data (training essays, features of test essays). Both the resulting prediction and variance estimates have closed-form solution, which helps speeding up the training and learning process substantially.
2. Empirical GPDW:

Parameter tuning of the scoring model is time consuming, which is also true for the GPDW being developed currently. We propose Empirical GPDW by estimating the hyperparameters using likelihood-based method, then fit those estimated hyperparameters into our model training. We hypothesize that this method saves us the effort of looking for the optimal hyperparameters through a grid search and may yield better performance.

1. Multivariate GPDW:

Both standard CRASE+ engine and GPDW being developed are trained seperately for each scoring dimension [footnote], which ignores the correlation among dimensions and hence leads to inefficient estimators in the model. Multivariate GPDW considers 4 dimensional scores simultaneously and takes the correlations among them into consideration during learning process.